

# Banzhaf Random Forests

Jianyuan Sun<sup>\*†</sup>, Guoqiang Zhong<sup>†\*\*\*\*</sup>, Junyu Dong, and Yajuan Cai

Department of Computer Science and Technology,  
Ocean University of China, Qingdao 266100, China  
sunjianyuan11@163.com, gqzhong@ouc.edu.cn, cyj-ouc@hotmail.com,  
dongjunyu@ouc.edu.cn

**Abstract.** Random forests are a type of ensemble method which makes predictions by combining the results of several *independent* trees. However, the theory of random forests has long been outpaced by their application. In this paper, we propose a novel random forests algorithm based on cooperative game theory. *Banzhaf power index* is employed to evaluate the power of each feature by traversing possible feature coalitions. Unlike the previously used information gain rate of information theory, which simply chooses the most informative feature, the Banzhaf power index can be considered as a metric of the importance of each feature on the *dependency* among a group of features. More importantly, we have proved the consistency of the proposed algorithm, named Banzhaf random forests (BRF). This theoretical analysis takes a step towards narrowing the gap between the theory and practice of random forests for classification problems. Experiments on several UCI benchmark data sets show that BRF is competitive with state-of-the-art classifiers and dramatically outperforms previous consistent random forests. Particularly, it is much more efficient than previous consistent random forests.

**Keywords:** random forests, Banzhaf power index, cooperative game, classification

## 1 Introduction

Ensemble methods are learning algorithms that construct a set of classifiers and combine them to classify new unseen data [1]. Random forests are a type of ensemble method based on combination of several independent decision trees [2]. In recent years, the random forests framework and its variants have been successfully applied in practice as a general classification and regression tool. Particularly, random forests have been widely used in computer vision [3], [4], [5], [6] and pattern recognition applications [7], [8], [9], [10], which promotes the state-of-the-art in performance. Despite their successful applications, the theoretical analysis of random forest models is still very difficult, even the basic mathematical properties are very hard to understand. In [11] and [12], Biau and colleagues tries to narrow the gap between the theory and practice

<sup>\*</sup> PhD Student

<sup>\*\*</sup> <sup>†</sup> Same contribution

<sup>\*\*\*</sup> To whom correspondence should be addressed.

of random forest. However, the proposed models in these two papers cannot deliver effective results and their running is not efficient.

In this paper, we introduce a novel random forests algorithm based on the cooperative game theory. We adopt the Banzhaf power index to evaluate the power of each feature by traversing all possible coalitions. Due to this, we call the proposed algorithm Banzhaf random forests (BRF). Different from the previously used information gain rate of information theory, which simply chooses the most informative feature, the Banzhaf power index measures the importance of each feature on the dependency among a group of features (coalition). More importantly, We reasonable proved the consistency of the forest, it has made a contribution to narrow the theory and practice gap for random classification forests problems.

The rest of this paper is organized as follows. In Section 2, we provide a brief overview of existing random forests models and analyze their advantage and disadvantage. In Section 3, we introduce the general random forests framework, including the construction of trees and randomness injection. Section 4 describes the proposed algorithm, Banzhaf random forests (BRF), in detail, while Section 5 is devoted to the justification of the consistency of BRF. Section 6 shows the experimental results on some UCI benchmark data sets and Section 7 concludes this paper.

## 2 Related work

Classic random forests introduced by Breiman [2] combine several decision trees [13] with bagging [14]. The main idea of random forests is based on the early work of [15] on the random subspace method, the feature selection work of [16], the way of random split selection of [17]. Based on the seminal work of Breiman [2], [18] suggests that it is best to average across sets of trees with different structures but not any of the constituent trees. Criminisi et al. [19] present a unified, efficient model of random decision forests which can be applied to a number of machine learning, computer vision and medical image analysis tasks. With the development of random forests in recent years, they have been applied to a wide variety of real world problems [20], [21], [22], [23].

Despite the successful applications of random forests in practice, the mathematical properties behind them have not been well understood. For example, the early theoretical work of [24], which is essentially based on mathematical heuristics, is not formalized to rigorous theory.

In theory, there are two main properties of theoretical interests related to random forests. One is the consistency of the models, that whether it can converge to an optimal solution as the data set grows infinitely large. The other is the rate of convergence. Our paper mainly focuses on consistency, which [11] has proved that Breiman's random forests cannot guarantee.

To design consistent random forests, many researchers have struggled in this trend. Meinshausen [25] has shown that an algorithm of random forests for quantile regression is consistent; Ishwaran and Kogalur [26] have shown the consistency of their survival forests model; Denil et al. [27] show the consistency of an online version of random forests, while [28] presents a new random regression forests. These consistent models can be applied to either regression, survival or online settings, but not to batch classi-

fication settings where all the training data can be used together for learning. In this paper, we propose a novel random forests model based on the cooperative game theory for multi-class classification problems. The consistency of the proposed algorithm is also proved.

Two more closely related papers to our work are [11] and [12]. [11] proves the consistency of some popular averaging classifiers, including random forests. Specifically, the authors take [2] as a weighted layered nearest neighbor classifier from the perspective of taxonomy proposed by [29]. Unfortunately, this property prevents the consistency of random tree classifiers. To remedy the inconsistency of tree classifiers, the authors suggest the technique introduced in [30]. Moreover, [11] has also proposed a scale-invariant version of random forests with consistency. Recently, [12] presents a new model of random forests, which is similar to the original algorithm of [2]. The main difference between these two models is in how random features are selected. [12] requires a second independent data set to evaluate the importance index of each feature and uses this property to prove the consistency for their algorithm, while the model of [2] doesn't need the second data set. In this paper, we use the Banzhaf power index to evaluate the power of each feature by traversing all possible feature coalitions, but not employing the second data set. The consistency of the proposed algorithm is theoretically guaranteed.

### 3 Random Forests

In this section we briefly review the random forests framework. Typically, random forests are built by combining the predictions of several trees, each of which is trained in isolation. Unlike in boosting [31], where the base models are trained and combined using a dynamic weighting scheme, the trees are trained independently and the predictions of the trees are combined through averaging or majority voting. For a more comprehensive review, please refer to [2] and [19].

To construct a random tree, three core steps are required: the first is the method for splitting the tree nodes; the second is the type of predictor to use in each leaf, and the third is the method of injecting randomness into the trees.

In a typical method for splitting nodes, splitting depends on whether or not they exceed a threshold value in a chosen feature. Alternatively, for linear splits, a linear combination of features are compared with a threshold to make decision. The threshold value in either case can be chosen randomly or by optimizing a function of the data. For example, the Gini index and information gain rate are commonly used. In this paper, we choose the midpoint of a feature as the splitting threshold, which leads to the proposed algorithm to be very efficient, especially in the case of large scale applications.

In order to split a node of each tree, candidate features of data are generated and a criterion is evaluated to choose between them. A simple strategy, as in the models analyzed in [11], is to choose among the features uniformly at random. A more common approach is to choose the candidate split which optimizes a purity function over the nodes that would be created. Particularly, two typical choices are to maximize the information gain [32] and minimize the Gini index. In our Banzhaf random forests, we

use the Banzhaf power index of the cooperative game theory [33], which measures the distribution of power among the features on the data sets.

For the choice of predictors, [19] propose several different leaf predictors for regression and other tasks. One common consideration is to average predictors over the training points which fall in that leaf. The other consideration may be based on majority voting with points in that leaf. In our work, we take the last strategy.

It is important to inject randomness into the trees for random forests. This can be achieved in several ways. One choice is on the features to be split at each node; the other one is the coefficients for random combinations of features. One common method is to build each tree using a bootstrapped or sub-sampled data set. In this way, each tree in the forest is trained on slightly different data, which introduces differences between the trees. Similar to [2], our work uses a bootstrapped method to inject randomness into each tree.

## 4 Banzhaf Random Forests

In this section, we describe the proposed algorithm, Banzhaf random forest (BRF), in detail. Firstly, we introduce some basic concepts of cooperative game theory. Secondly, based on the Banzhaf power index, we introduce the way to construct the randomized trees. Thirdly, we combine the Banzhaf trees to formulate the Banzhaf random forests. Finally, we present the prediction method about the Banzhaf random forests.

### 4.1 Basic concepts of cooperative game theory

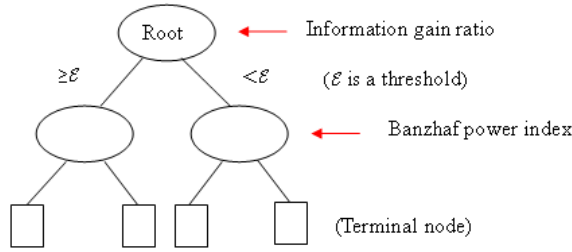
Cooperative game theory mainly studies an ‘acceptable’ way of distributing gains collectively achieved by a group of cooperating agents [34]. A cooperative profit game  $\Gamma = (\mathcal{N}, \gamma)$  consists of a player set  $\mathcal{N} = \{1, 2, \dots, n\}$  and a characteristic function  $\gamma : 2^{\mathcal{N}} \rightarrow R$ . For each subset  $S \subseteq \mathcal{N}$ ,  $\gamma(S)$  can be interpreted as the profit achieved by the players in  $S \subseteq \mathcal{N}$ . The usual goal in cooperative game is to distribute the total gain  $\gamma(\mathcal{N})$  of the global coalition  $\mathcal{N}$  among each player in fair and reasonable ways. Different requirements on the fairness and rationality derive different solution concepts of the cooperative game. Such as the core, the Banzhaf power index and some related concepts of approximate core. Among various solution concepts the concept of Banzhaf power index that is motivated by fairness.

For a game  $\Gamma = (\mathcal{N}, \gamma)$ , if it is monotone, i.e., it satisfies  $\gamma(\mathcal{C}) \leq \gamma(\mathcal{D})$  for every pair of coalitions  $\mathcal{C}, \mathcal{D} \subseteq \mathcal{N}$  such that  $\mathcal{C} \subseteq \mathcal{D}$ , and its characteristic function only takes value 0 and 1, i.e.,  $\gamma(S) \in \{0, 1\}$ ,  $\forall S \subseteq \mathcal{N}$ , this game is called a simple game. In a simple game  $\Gamma = (\mathcal{N}, \gamma)$ , the coalitions with value 1 are called ‘winning’, and that with value 0 are called ‘losing’, i.e.,  $\forall S \subseteq \mathcal{N}$ ,  $\gamma(S) = 1$  and  $\gamma(S) = 0$ , respectively. Each coalition  $S \cup \{i\}$  that wins when  $S$  loses is called a swing for player  $i \in \mathcal{N}$ , because the membership of player  $i$  in the coalition is crucial to the ‘winning’. In fact, Banzhaf power index is to count the number of winning coalitions, when the player  $\forall i \in \mathcal{N}$  joining some losing coalitions, to find the most crucial player that it can let the majority of coalitions winning.

Banzhaf power index, which yields a unique outcome in coalitional games, is proposed to measure the marginal contribution of players in the game [33]. In simple games, the Banzhaf power index has a particular attractive interpretation: it measures the power of a player, i.e., the probability that he can influence the outcome of the game. In this paper, we use Banzhaf power index to measure the power of each feature.

## 4.2 Construction of Banzhaf tree

Figure 1 shows the structure of a Banzhaf decision tree. For the root node, the feature is selected with information gain rate. For all the other nodes, the features are selected with the Banzhaf power index. The idea of Banzhaf decision tree are mainly motivated by game theory, especially, the cooperative game theory. We take the features of data as the players in a game, then the original tree construction problem is transformed into a cooperative ‘feature’ game. At each node, features in the form of the coalition are selected and the best one is split.



**Fig. 1.** A Banzhaf tree

Next, we first present the way to compute the Banzhaf power index in this work.

The original definition of Banzhaf power index is described in [33]. Given a cooperative game  $\Gamma = (\mathcal{N}, \gamma)$  with  $|\mathcal{N}| = n$ , the Banzhaf power index of a player  $i \in \mathcal{N}$  is the probability of swings for play  $i$ . We denote the Banzhaf power index as  $\beta_i(\Gamma)$  and it is given by

$$\beta_i(\Gamma) = \frac{1}{|2^{\mathcal{N} \setminus i}|} \sum_{S \subseteq \mathcal{N}} \Delta_i(S), \quad (1)$$

where  $\Delta_i(S)$  is the marginal contribution of player  $i$ . i.e.  $\Delta_i(S) = \gamma(S \cup i) - \gamma(S)$ .

Banzhaf power index measures the distribution of power among the players in cooperative games. Here, we apply it for the decision tree construction, attempting to estimate the power of each feature for each tree node. The power of each feature can be measured by averaging the contributions that it makes to each of the subset which it belongs to. Let coalition  $\mathcal{K}$  be a candidate feature subset and feature  $f_i (f_i \notin \mathcal{K})$  is to be estimated. Define the ratio  $p = \mu_i(\mathcal{K}) / \rho_i(\mathcal{K})$  to represent the impact of feature  $f_i$  on coalition  $\mathcal{K}$ , where  $\mu_i(\mathcal{K})$  can be interpreted as the number of features that fall

into interdependence relationship with the feature  $f_i$ , and  $\rho_i(\mathcal{K})$  be the number of features in the coalition  $\mathcal{K}$ . Therefore, we define a threshold value  $\tau$ . If  $p < \tau$  (commonly  $\tau = 1/2$ ), we call the coalition  $\mathcal{K} \cup f_i$  ‘losing’, otherwise ‘winning’, i.e.

$$\Delta_i(\mathcal{K} \cup f_i) = \begin{cases} 1 & p \geq \tau; \\ 0 & p < \tau. \end{cases} \quad (2)$$

Here,  $\Delta_i(\mathcal{K} \cup f_i) = 1$  means that feature  $f_i$  is the key to make the coalition to exhibit better performance. The threshold value  $1/2$  means, if more than half of the features are interdependent with  $f_i$ , it will join in the coalition to make it ‘winning’. Hence, for simplicity of the computation, we define  $\Delta_i(\mathcal{S})$  in Eq. (1) as

$$\Delta_i(\mathcal{S}) = \begin{cases} 1 & p \geq \tau; \\ 0 & p < \tau. \end{cases} \quad (3)$$

For clarity, here, we give an example to show how to compute the Banzhaf power index. Given a cooperative ‘feature’ game  $\Gamma = (\mathcal{N}, \gamma)$ , the feature player set  $\mathcal{N} = \{f_1, f_2, f_3, f_4\}$ . Suppose, currently, the goal is to calculate the Banzhaf power index of  $f_4$ . The total number of possible coalitions of feature subsets  $\mathcal{N} \setminus f_4$  is 7 (except  $\emptyset$ ), for all  $\mathcal{S} \subseteq \mathcal{N} \setminus f_4$ . Assume the winning coalitions with respect to  $f_4$  are  $\{f_2\}$ ,  $\{f_2, f_3\}$ ,  $\{f_1, f_2\}$ , i.e. half of the coalitions are interdependent with feature  $f_4$ . Then the Banzhaf power index of  $f_4$  can be computed as

$$\beta_i(\Gamma) = \frac{1}{|2^{\mathcal{N} \setminus f_4}|} \sum_{\mathcal{S} \subseteq \mathcal{N}} \Delta_i(\mathcal{S}) = 3/7. \quad (4)$$

Similarly, the value of Banzhaf power index for other features can be computed as the same way. Generally, Banzhaf power index is hardly to be zero in large scale and high dimensional applications.

In order to evaluate the impact of feature  $f_i$ , it needs to calculate the proportion of the ‘winning’ coalitions. That will lead to a high computational complexity, but our model only randomly selected a small group of features to compute the Banzhaf power index at each node. Hence, the computational complexity is fairly low.

To calculate the proportion of the ‘winning’ coalitions, we use conditional mutual information of information theory to evaluate the interdependent between a single  $f_j \notin \mathcal{S} \subseteq \mathcal{N}$  and the feature player  $f_i \in \mathcal{S} \subseteq \mathcal{N}$ . If more than half of feature players  $f_i \in \mathcal{S}$  are interdependent  $f_j$ , then have  $\Delta_j(\mathcal{S}) = \gamma(\mathcal{S} \cup j) - \gamma(\mathcal{S}) = 1$ .

In our paper, the condition mutual information is defined as the amount of the interdependent between feature player  $f_j \notin \mathcal{S}$  and feature player  $f_i \in \mathcal{S}$  given the feature player colation  $\mathcal{S}$ . It is formally defined by

$$I(f_j; f_i | \mathcal{S} \setminus f_i) = \sum_{x \in f_j} \sum_{y \in f_i} \sum_{z \in \mathcal{S} \setminus f_i} \log \frac{p(x, y | z)}{p(x | z)p(y | z)}. \quad (5)$$

By Eq. (1), (3) and (5), we can get the Banzhaf power index of each feature player for the construction of each decision tree.

### 4.3 Banzhaf random forests algorithm

Given a training data set  $D_n = (X_i, Y_i)_{i=1}^n$ , it includes  $n$  samples and the dimensionality of data is  $M$ . The procedures of the Banzhaf random forests (BRF) algorithm can be described as follows.

- For the construction of each Banzhaf decision tree in BRF, randomly draw  $n$  samples with replacement using bootstrap and randomly select  $h \ll M$  features without replacement from the training data. Base on this data set  $d_n = (X_i, Y_i)_{n \times (h+1)} \subseteq D_n = (X_i, Y_i)_{n \times (h+1)}$ , grow a recursive Banzhaf tree.
- For the root node, the feature is selected with information gain rate. For all the other nodes, the features are selected with the Banzhaf power index. The feature associated with the corresponding node is split at the midpoint of the feature values, to generate the left and right branches.
- If a (terminal) node has the percentage of incorrectly assigned samples less than  $d$ , then stop building the Banzhaf tree, where  $d$  is a pre-specified number.
- BRF predicts the labels of test data based on the votes it received from each Banzhaf tree.

Our algorithm is similar to the original algorithms of [2]. Both of them used bootstrap aggregating i.e., bagging ensemble algorithm. The main difference between BRF and the algorithm of [2] is in how the feature associated with a node is selected. BRF uses Banzhaf power index, while Breiman's method use the Gini index. Another difference is, BRF splits each node at the midpoint of the feature values but Breiman's algorithm does not. More importantly, as shown in next section, the consistency of BRF is theoretically guaranteed, but that of Breiman's algorithm is not.

We have also tested the model of pure Banzhaf random forests, i.e. the feature of the root node is also selected via the Banzhaf power index. Their performance is generally worse than that of the BRF algorithm described as above. One reason for this result may be that the feature selected via information gain rate at the root node may present some important invariant information of data.

### 4.4 Prediction

We denote a recursive tree created in the BRF algorithm based on data  $D_n = (X_i, Y_i)_{i=1}^n$  as  $g_n$ , where  $(X_i, Y_i)_{i=1}^n$  are i.i.d. pairs of random variables such that  $X$  (the feature vector) takes its value in  $R^d$  while  $Y$  (the label) is a multiclass random variable. To make a prediction for a query point  $x$ , each Banzhaf decision tree computes,

$$\zeta_n^k(x) = \frac{1}{N(A_n(x))} \sum_{(X_i, Y_i) \in A_n(x)} \delta(Y_i = k),$$

where  $A_n(x)$  denotes the node of the tree containing  $x$ , and  $N(A_n(x))$  is the number of points that located in  $A(x)$ . Then the tree prediction is the class which maximizes that:

$$g_n(x) = \arg \max_k \{\zeta_n^k(x)\}.$$

The forest predicts the class with the most votes from the individual trees.

## 5 Consistency

In this section, we prove the consistency of Banzhaf random forests. We denote the Banzhaf tree created by Banzhaf random forests trained on data  $(X_i, Y_i)_{i=1}^n$  as  $g_n$ . The consistency of a sequence  $\{g_n\}$  is defined as follows.

**Definition 1** A sequence of classifier  $\{g_n\}$  is consistent for a given distribution of  $(X, Y)$ , that is, the probability of prediction error of  $g_n$  converges in probability to the Bayesian risk,

$$L(g_n) = \mathbb{P}(g_n(X, \theta) \neq Y | D_n) \rightarrow L^*,$$

as  $n \rightarrow \infty$ . Here,  $\theta$  denotes the randomness in the tree-building algorithm,  $D_n$  is the training data set and the probability in the convergence is over the random selection of  $D_n$ . The Bayesian risk is the probability of prediction error of the Bayesian classifier, which makes predictions by choosing the class with the highest posterior probability,  $g(x) = \arg \max_k \mathbb{P}(Y = k | X = x)$ .

In order to reduce the complexity of the issue, we consider that multi-class classifier can be transformed to combination of several binary-class classifier. So, we need to prove the consistency of estimators of the posterior distribution of each class. A similar result was shown by Denil et al [27].

**Lemma 1** Suppose we have the estimates,  $\zeta_n^k(x)$ , for each class posterior  $\zeta^k(x) = \mathbb{P}(Y = k | X = x)$  and that these estimates are each consistent. The classifier

$$g_n(x) = \arg \max_k \{\zeta_n^k(x)\}$$

is consistent for the corresponding multi-class classification problem.

Proof. By definition, the rule

$$g(x) = \arg \max_k \{\zeta^k(x)\}$$

achieves the Bayes risk. In the case where all the  $\zeta^k(x)$  are equal there is nothing to prove, since all choices have the same probability of error. So, suppose there is at least one  $k$  such that  $\zeta^k(x) < \zeta^{g(x)}(x)$  and define

$$\begin{aligned} m(x) &= \zeta^{g(x)}(x) - \max_k \{\zeta^k(x) | \zeta^k(x) < \zeta^{g(x)}(x)\} \\ m_n(x) &= \zeta_n^{g(x)}(x) - \max_k \{\zeta_n^k(x) | \zeta_n^k(x) < \zeta_n^{g(x)}(x)\} \end{aligned}$$

The function  $m(x) \geq 0$  is the margin function which measures how much better the best choice is than the second best choice. The function  $m_n(x)$  measures the margin of  $g_n(x)$ . If  $m_n(x) > 0$  then  $g_n(x)$  has the same probability of error as the Bayes classifier.

The assumption above guarantees that there is some  $\epsilon$  such that  $m(x) > \epsilon$ . Using  $\mathcal{C}$  to denote the number of classes, by making  $n$  large it can satisfy

$$\mathbb{P}(|\zeta_n^k(X) - \zeta^k(X)| < \epsilon/2) \geq 1 - \delta/\mathcal{C}$$



since  $\zeta_n^k$  is consistent. Thus

$$\mathbb{P}\left(\bigcap_{k=1}^C |\zeta_n^k(X) - \zeta^k(X)| < \epsilon/2\right) \geq 1 - K + \sum_{k=1}^C \mathbb{P}(|\zeta_n^k(X) - \zeta^k(X)| < \epsilon/2) \geq 1 - \delta$$

So with probability at least  $1 - \delta$  we have

$$\begin{aligned} m_n(X) &= \zeta_n^{g(X)} - \max_k \{\zeta_n^k(X) | \zeta^k(X) < \zeta^{g(X)}(X)\} \\ &\geq (\zeta^{g(X)} - \epsilon/2) - \max_k \{\zeta_n^k(X) + \epsilon/2 | \zeta^k(X) < \zeta^{g(X)}(X)\} \\ &= \zeta^{g(X)} - \max_k \{\zeta^k(X) | \zeta^k(X) < \zeta^{g(X)}(X)\} - \epsilon > 0 \end{aligned}$$

Since  $\delta$  is arbitrary this means that the risk of  $g_n$  converges in probability to the Bayes risk.

Lemma 1 allows us to prove the consistency of the multiclass classifier can be transformed to prove the consistency of several two class posterior estimates. i.e., Given a set of classes  $\{1, \dots, c\}$  we can re-assign the labels using the map  $(X, Y) \mapsto (X, \mathcal{I}(Y = k))$  for any  $k \in \{1, \dots, c\}$  in order to get a two class problem where  $\mathbb{P}(Y = 1 | X = x)$  in this new problem is equal to  $\zeta^k(x)$  in the original multiclass problem.

Then, we are inspired by [27]. The following Lemma 2 allows us to focus our attention on the consistency of each of the tree estimators in the classification forests.

**Lemma 2** Assume that the sequence  $\{g_n\}$  of randomized classifiers is consistent for a certain distribution of  $(X, Y)$ . Then the voting classifier  $g_n^{(m)}$  obtained by taking the majority vote over  $M$  (not necessarily independent) copies of  $\{g_n\}$  is also consistent.

Proof. Let  $g(x)$  denote the Bayes classifier. Consistency of  $\{g_n\}$  is equivalent to saying that  $\mathbb{E}[L(g_n)] = \mathbb{P}(g_n(X, \theta) \neq Y) \rightarrow L^*$ . In fact, since  $\mathbb{P}(g_n(X, \theta) \neq Y | X = x) \geq \mathbb{P}(g(X) \neq Y | X = x)$  for all  $x \in \mathbb{R}^D$ , consistency of  $\{g_n\}$  means that  $\mu$ -almost all  $x$ ,

$$\mathbb{P}(g_n(X, \theta) \neq Y | X = x) \rightarrow \mathbb{P}(g(X) \neq Y | X = x) = 1 - \max_k \{\zeta^k(x)\}$$

Define the following indices

$$G = \{k | \zeta^k(x) = \max_k \{\zeta^k(x)\}, B = \{k | \zeta^k(x) < \max_k \{\zeta^k(x)\}\}$$

Then

$$\begin{aligned} \mathbb{P}(g_n(X, \theta) \neq Y | X = x) &= \sum_k \mathbb{P}(g_n(X, \theta) = k | X = x) \mathbb{P}(Y \neq k | X = x) \\ &\leq (1 - \max_k \{\zeta^k(x)\}) \sum_{k \in G} \mathbb{P}(g_n(X, \theta) = k | X = x) + \sum_{k \in B} \mathbb{P}(g_n(X, \theta) = k | X = x) \end{aligned}$$

which means it suffices to show that  $\mathbb{P}(g_n^{(m)}(X, \theta^M) = k | X = x) \rightarrow 0$  for all  $k \in B$ . However, using  $\theta^M$  to denote  $M$  (possibly dependent) copies of  $\theta$ , for all  $k \in B$  we

have

$$\begin{aligned}\mathbb{P}(g_n^{(m)}(X, \theta^M) = k) &= \mathbb{P}\left(\sum_{j=1}^M \mathbb{I}\{g_n(x, \theta_j) = k\} > \max_{c \neq k} \sum_{j=1}^M \mathbb{I}\{g_n(x, \theta_j) = c\}\right) \\ &\leq \mathbb{P}\left(\sum_{j=1}^M \mathbb{I}\{g_n(x, \theta_j) = k\} \geq 1\right)\end{aligned}$$

By Markov's inequality,

$$\begin{aligned}&\leq \mathbb{E}\left[\sum_{j=1}^M \mathbb{I}\{g_n(X, \theta_j) = k\}\right] \\ &= M\mathbb{P}(g_n(X, \theta) = k) \rightarrow 0.\end{aligned}$$

According to Lemma 2, we conclude that the consistency of Banzhaf random forests is implied by the consistency of the trees which composed of. In addition, we use the bagging ensemble method to construct BRF. So by the Theorem 1 in [11], we know that the consistency of a voting Banzhaf random forests which follows from the consistency of the base classifier. Here, Biau et al. introduce a parameter  $q_n \in [0, 1]$ . In the bootstrap sample  $D_n(\theta)$ , each data pair  $(X_i, Y_i)$  is present with probability  $q_n$  which is independent from each other.

**Theorem 1** Let  $\{g_n\}$  be a sequence of classifier that is consistency for the distribution of  $(X, Y)$ . Consider the Banzhaf random forests (majority voting classifiers)  $g_n^{(m)}(X, \theta^m, D_n)$ , using parameter  $q_n$ . If  $nq_n \rightarrow \infty$  as  $n \rightarrow \infty$  then both classifiers are consistent.

Proof. See that for Theorem 1 in [11].

With Lemma 2 and Theorem 1 established, the remainder of effort goes into proving the consistency of a Banzhaf tree construction. For each tree in the Banzhaf forests is established based on the Banzhaf index. We show that if a classifier is condition consistency which consists of a small group of random variable, and uses the Banzhaf power index to sampling for this sample process for this random variable generates acceptable sequences with probability 1, then the resulting classifier is unconditionally consistent.

**Theorem 2** Suppose  $\{g_n\}$  is a sequence of classifiers whose probability of error converges conditionally in probability to the Bayes risk  $L^*$  for a specified distribution on  $(X, Y)$ , i.e.

$$\mathbb{P}(g_n(X, \theta, I) \neq Y | I) \rightarrow L^*,$$

for all  $I \in \mathcal{I}$ ,  $I$  is a random sequence produced by Banzhaf power index, and that  $v$  is a distribution on  $I$ . If  $v(\mathcal{I}) = 1$  which means produce acceptable sequence with probability value is 1, then the probability of error converges unconditionally in probability, i.e.

$$\mathbb{P}(g_n(X, \theta, I) \neq Y) \rightarrow L^*,$$

$\{g_n\}$  is consistent for the specified distribution.

Proof. The sequence in question is uniformly integrable, so it is sufficient to show that  $\mathbb{E}[\mathbb{P}(g_n(X, \theta, I) \neq Y|I)] \rightarrow L^*$  implies the result, where the expectation is taken over the random selection of training set and  $I$  is the specific structure of the tree,  $\{g_n\}$ . We can write

$$\begin{aligned}\mathbb{P}(g_n(X, \theta, I) \neq Y) &= \mathbb{E}[\mathbb{P}(g_n(X, \theta, I) \neq Y|I)] \\ &= \int_{\mathcal{I}} \mathbb{P}(g_n(X, \theta, I) \neq Y|I)v(I) + \int_{\mathcal{I}^c} \mathbb{P}(g_n(X, \theta, I) \neq Y|I)v(I)\end{aligned}$$

By assumption  $v(\mathcal{I}^c) = 0$  then we have

$$\lim_{n \rightarrow \infty} \mathbb{P}(g_n(X, \theta, I) \neq Y) = \lim_{n \rightarrow \infty} \int_{\mathcal{I}} \mathbb{P}(g_n(X, \theta, I) \neq Y|I)v(I)$$

Since probabilities are bounded in the interval  $[0, 1]$ , the dominated convergence theorem allows us to exchange the integral and the limit,

$$= \int_{\mathcal{I}} \lim_{n \rightarrow \infty} \mathbb{P}(g_n(X, \theta, I) \neq Y|I)v(I)$$

and by assumption the conditional risk converges to the Bayes risk for all  $I \in \mathcal{I}$ , so

$$= L^* \int_{\mathcal{I}} v(I) = L^*$$

which is the desired result.

In fact, let the Banzhaf power index  $\eta(f_i)$  is equal to the income distribution function  $\gamma(f_i)$  in a tree construction game  $\Gamma = (\mathcal{N}, \gamma)$ , i.e.,  $\eta(f_i) = \gamma(f_i)$ . Because we chose the maximize Banzhaf power index for each node of each tree. We can obtain a acceptable random variable sequence that all with the maximize Banzhaf power index. By  $\eta(f_i) = \gamma(f_i)$ , these random variable sequence cooperative can obtain the best result. So it is sufficient to show that the Banzhaf tree is consistent conditioned on such a sequence.

In conclusion, we proved the consistency of our tree construct by the Theorem 2. Because the Theorem 1 is established, we can achieve the consistency of Banzhaf random forests.

## 6 Experiments

To evaluate the proposed algorithm, BRF, we tested it on several data sets from the UCI machine learning repository, including iris, wine, ecoli, thyroid, soybean, shuttle, dermatology, sonar and musk2. We compare it with Breiman's random forests [2] and the model proposed in [12]. We implemented Breiman's random forest with C4.5 as it generally performs well on classification problems. As mentioned above, the model proposed in [12] is consistent. For comparison, we also listed the classification results yielded by k-nearest neighbor classifier (KNNs) and support vector machines (SVM).

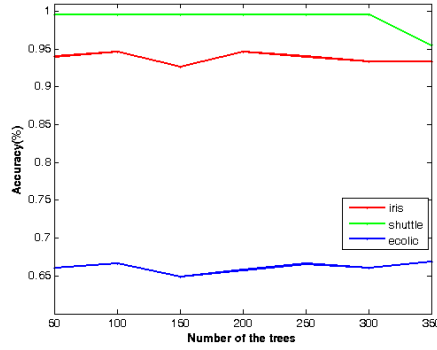
Table 1 shows the specific information of the used UCI data sets.

Datasets	No.examples	No.features	No.classes
soybean	47	35	4
iris	150	4	3
wine	178	13	3
sonar	208	20	2
thyroid	215	5	3
ecoli	357	7	8
dermatology	366	34	6
musk2	6598	166	2
shuttle	14516	9	7

**Table 1.** Summary of the used UCI data sets.

### 6.1 Effect of the number of trees in BRF

To evaluate the effect of the number of trees in BRF, we conducted experiments on three data sets: iris, ecoli and shuttle. Fig. 2 shows the obtained classification accuracy against the number of trees in BRF. We can see that, BRF is basically robust with the number of trees. Particularly, when the number of trees equals to 100, BRF performs slightly better than other values.

**Fig. 2.** Effect of the number of trees in BRF.

### 6.2 Comparison on running efficiency

To test the running speed of BRF, we performed experiments on seven data sets: iris, wine, ecoli, soybean, thyroid, dermatology and shuttle. We compared it with the model of [2] and that of [12]. From Table 2, we can see that, the running of BRF is slower than the model of [2]. This is mainly because calculation of the Banzhaf power index needs some time when constructing the trees. However, BRF is more efficient than the model of [12], which is a state-of-the-art consistent random forests model.

Datasets	Breiman01	Biau12	BRF
iris	1.321	3.107	1.654
wine	5.401	16.781	9.134
ecoli	5.729	17.438	8.778
soybean	0.673	5.761	2.297
thyroid	2.857	4.856	3.168
dermatology	2.463	71.201	11.023
shuttle	49.71	39600.63	80.660

**Table 2.** Running time of two compared models and BRF on seven UCI data sets (the unit is second).

### 6.3 Classification results

To evaluate BRF on multi-class classification problems, we compared it with KNNs, SVMs, the model of [2], and the model of [12]. Nine UCI data sets were used. They are iris, wine, ecoli, thyroid, soybean, shuttle, dermatology, sonar and musk2. For all these data sets, we used 5-fold cross validation to test the models. The average classification accuracies are reported. For the model of [2] and BRF, we used the same number of trees in the random features. Following Breiman's suggestion for classification problems [2], we set the number of trees to  $\text{round}(\log 2(h) + 1)$ , where  $h$  is the dimensionality of features. To be fair, we set up the same termination conditions for all the random forests models, i.e. the percentage of incorrectly assigned samples at the termination node should be no greater than the number of classes on a data set. For KNNs and SVMs, we selected the parameter with 5-fold cross validation.

Table 3 shows the results obtained by the compared models and BRF. We can see that BRF performs slightly better than KNNs, SVMs and the model of [2], and consistently better than the model of [12]. This demonstrates that using interdependent features to construct the randomized trees can lead to better results than using independent features in random forests.

Datasets	KNN	SVM	Breiman01	Biau12	BRF
soybean	<b>1.0000</b>	<b>1.0000</b>	<b>1.0000</b>	0.5717	<b>1.0000</b>
iris	0.9467	<b>0.9867</b>	0.9467	0.8353	0.9467
wine	0.9423	0.6782	0.9599	0.5580	<b>0.9717</b>
sonar	0.5908	0.6583	0.7032	0.5819	<b>0.7120</b>
thyroid	0.9395	0.9023	<b>0.9488</b>	0.8000	0.9395
ecoli	0.8356	<b>0.8431</b>	0.5958	0.4286	0.6665
dermatology	0.9656	0.9540	0.9589	0.4397	<b>0.9677</b>
musk2	0.7227	0.8508	0.8509	0.6542	<b>0.8710</b>
shuttle	0.9951	0.9752	0.9957	0.8256	<b>0.9957</b>

**Table 3.** Classification accuracy obtained by the compared models and BRF on the UCI data sets.

## 7 Conclusion

In this paper, we propose a novel random forests model called Banzhaf random forests (BRF) based on the concepts of the cooperative game theory. It's consistency is proved, which takes a step towards narrowing the gap between the theory and practice of random forest. This work is probably the first one that apply the cooperative game theory to random forests, and we have tested and verified the feasibility of the idea. Experiments on UCI data sets show that BRF not only slightly outperforms state-of-the-art classifiers, including KNNs, SVMs and the random forests model by Breiman [2], but much more efficient than existing consistent random forests.

## Acknowledgment

This research was supported by the National Natural Science Foundation of China (NSFC) under Grant no. 61271405 and 61403353, and the Fundamental Research Funds for the Central Universities of China.

## References

1. Zhou, Zhi-Hua: Ensemble methods: foundations and algorithms. CRC Press (2012)
2. Breiman, Leo.: Random forests. Machine learning, vol. 45, pp. 5–32. Springer (2001)
3. Lepetit, Vincent and Fua, Pascal: Keypoint recognition using randomized trees. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 28, pp. 1465–1479. IEEE (2006)
4. Ozuysal, Mustafa and Fua, Pascal and Lepetit, Vincent: Fast keypoint recognition in ten lines of code. Computer Vision and Pattern Recognition, 2007, CVPR'07. pp. 1–8. Ieee (2007)
5. Shotton, Jamie and Sharp, Toby and Kipman, Alex and Fitzgibbon, Andrew and Finocchio, Mark and Blake, Andrew and Cook, Mat and Moore, Richard: Real-time human pose recognition in parts from single depth images. Communications of the ACM, vol. 56, pp. 116–124. ACM (2013)
6. Zikic, Darko and Glocker, Ben and Criminisi, Antonio: Atlas encoding by randomized forests for efficient label propagation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013, pp. 66–73. Springer (2013)
7. Winn, John and Criminisi, Antonio: Object class recognition at a glance. In Video Proc. CVPR (2006)
8. Yin, Pei and Criminisi, Antonio and Winn, John and Essa, Irfan: Tree-based classifiers for bi-layer video segmentation. Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pp. 1–8. IEEE (2007)
9. Bosch, Anna and Zisserman, Andrew and Muoz, Xavier: Image classification using random forests and ferns. Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pp. 1–8. IEEE (2007)
10. Shotton, Jamie and Johnson, Matthew and Cipolla, Roberto: Semantic texton forests for image categorization and segmentation. Computer vision and pattern recognition, 2008. CVPR 2008. IEEE Conference on, pp. 1–8. IEEE (2008)
11. Biau, Gérard and Devroye, Luc and Lugosi, Gábor: Consistency of random forests and other averaging classifiers. The Journal of Machine Learning Research, vol. 9, pp. 2015–2033. JMLR. org (2008)

12. Biau, Gérard: Analysis of a random forests model. *The Journal of Machine Learning Research*, vol. 13, pp. 1063–1095. JMLR. org (2012)
13. Breiman, Leo and Friedman, Jerome and Stone, Charles J and Olshen, Richard A.: *Classification and regression trees*. CRC press (1984)
14. Breiman, Leo.: Bagging predictors. *Machine learning*, vol. 24, pp. 123–140. Springer (1996)
15. Ho, Tin Kam: The random subspace method for constructing decision forests, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, pp. 832–844. IEEE (1998)
16. Amit, Yali and Geman, Donald: Shape quantization and recognition with randomized trees. *Neural computation*, vol. 9, pp. 1545–1588. MIT Press (1997)
17. Dietterich, Thomas G.: An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, vol. 40, pp. 139–157. Springer (2000)
18. Kwok, Suk Wah and Carter, Chris: Multiple decision trees. *arXiv preprint arXiv:1304.2363* (2013)
19. Criminisi, Antonio and Shotton, Jamie and Konukoglu, Ender: Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision*, pp. 81–227 (2012)
20. Svetnik, Vladimir and Liaw, Andy and Tong, Christopher and Culberson, J Christopher and Sheridan, Robert P and Feuston, Bradley P.: Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, vol. 43, pp. 1947–1958. ACS Publications (2003)
21. Prasad, Anantha M and Iverson, Louis R and Liaw, Andy: Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, vol. 9, pp. 181–199. Springer (2006)
22. Cutler, D Richard and Edwards Jr, Thomas C and Beard, Karen H and Cutler, Adele and Hess, Kyle T and Gibson, Jacob and Lawler, Joshua J.: Random forests for classification in ecology. *Ecology*, vol. 88, pp. 2783–2792. *Eco Soc America* (2007)
23. Criminisi, Antonio and Shotton, Jamie: *Decision forests for computer vision and medical image analysis*. Springer Science & Business Media (2013)
24. Breiman, Leo.: Consistency for a simple model of random forests. *Statistical Department, University of California at Berkeley. Technical Report*, (2004)
25. Meinshausen, Nicolai: Quantile regression forests. *The Journal of Machine Learning Research*, vol. 7, pp. 983–999. JMLR. org (2006)
26. Ishwaran, Hemant and Kogalur, Udaya B.: Consistency of random survival forests. *Statistics & probability letters*, vol. 80, pp. 1056–1064. Elsevier (2010)
27. Denil, Misha and Matheson, David and de Freitas, Nando: Consistency of online random forests. *arXiv preprint arXiv:1302.4853* (2013)
28. Denil, Misha and Matheson, David and De Freitas, Nando: Narrowing the gap: Random forests in theory and in practice. *arXiv preprint arXiv:1310.1415*, (2013)
29. Lin, Yi and Jeon, Yongho: Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, vol. 101, pp. 578–590. Taylor & Francis (2006)
30. Györfi, L and Devroye, L and Lugosi, G.: *A probabilistic theory of pattern recognition*. Springer-Verlag, (1996)
31. Schapire, Robert E and Freund, Yoav: *Boosting: Foundations and Algorithms*. Kybernetes, vol. 42, pp. 164–166. Emerald Group Publishing Limited (2013)
32. Hastie, Trevor and Tibshirani, Robert and Friedman, Jerome and Hastie, T and Friedman, J and Tibshirani, R.: *The elements of statistical learning*. vol. 2. Springer(2009)
33. Banzhaf III, John F.: Weighted voting doesn't work: A mathematical analysis. *Rutgers L. Rev.*, vol. 19, pp. 317. HeinOnline (1964)

34. Chalkiadakis, Georgios and Elkind, Edith and Wooldridge, Michael: Computational aspects of cooperative game theory. Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 5, pp. 1–168. Morgan & Claypool Publishers (2011)